Capstone 2:

Detecting Deceptive Hotel Reviews

Through Topic Modeling and Machine Learning

SpringBoard Data Science Career Program

Chris Gian

github.com/chrisgian

Roadmap

- Motivation
- Data
- Data Preparation
- Topic Modeling
- Classification and Tuning
- Results
- Future steps

Motivation (1)

Deceptive text-based content plagues many industries and aspects of our information-rich lives.

1. Yelp Fake reviews
   - Increase a business' popularity in rankings
2. Spam
   - Unwanted messages show up on cell phones, emails, and social media
3. Fake news
   - Biased and false content published under the guise of unbiased and legitimate news

Motivation (2)

Yelp's fake review problem

Daniel Roberts
Sep 26, 2013

FORTUNE -- On Monday, New York State Attorney General Eric Schneiderman announced the conclusion of "Operation Clean Turf," a yearlong sting that caught 19 different companies, most of them SEO (search engine optimization) or reputation management firms for hire, that were writing fake reviews for small businesses that paid them. The NYAG fined the businesses in varying amounts that total $350,000. (As part of the operation, people from the NYAG's office even posed as owners of a Brooklyn yogurt shop.) In the press release, Schneiderman says that the investigation "tells us that we should approach online reviews with caution" and calls the process of posting fake reviews online, "the 21st century's version of false advertising."

Motivation (3) Approach

This capstone project focuses on using **Natural Language Processing** and **Machine Learning** methods to identify deceptive content.

Motivation (4) Hypothetical Client

A hotel booking aggregator is trying to improve the quality and reliability of its hotel booking information for its users.

- It has noticed complaints after users voiced concerns that some reviews may be fake.
- To alleviate this, the client seeks to build a predictive model to flag whether reviews are genuine or fake.
- By the end of this project, the client will be able to use this algorithm as part of their fake review deterrence strategy.

Data (1) Overview

Source: Kaggle, “Deceptive Hotel Opinions Corpus”

- 1600 records
  - 800 Deceptive
    - Generated by Mechanical Turks: Crowdsourced, human intelligence tasking
  - 800 Genuine
    - Multiple Review Aggregator sites
    - Web
  - Positive and Negative Sentiment

Data (2) Descriptives

Top 5 and Bottom 5 Records

	deceptive	hotel	polarity	source	text
0	truthful	conrad	positive	TripAdvisor	We stayed for a one night getaway with family ...
1	truthful	hyatt	positive	TripAdvisor	Triple A rate with upgrade to view room was le...
2	truthful	hyatt	positive	TripAdvisor	This comes a little late as I'm finally catchi...
3	truthful	omni	positive	TripAdvisor	The Omni Chicago really delivers on all fronts...
4	truthful	hyatt	positive	TripAdvisor	I asked for a high floor away from the elevato...
1595	deceptive	intercontinental	negative	MTurk	Problems started when I booked the InterContin...
1596	deceptive	amalfi	negative	MTurk	The Amalfi Hotel has a beautiful website and i...
1597	deceptive	intercontinental	negative	MTurk	The Intercontinental Chicago Magnificent Mile ...
1598	deceptive	palmer	negative	MTurk	The Palmer House Hilton, while it looks good i...
1599	deceptive	amalfi	negative	MTurk	As a former Chicagoan, I'm appalled at the Ama...

Positive and Negative Sentiment by Source

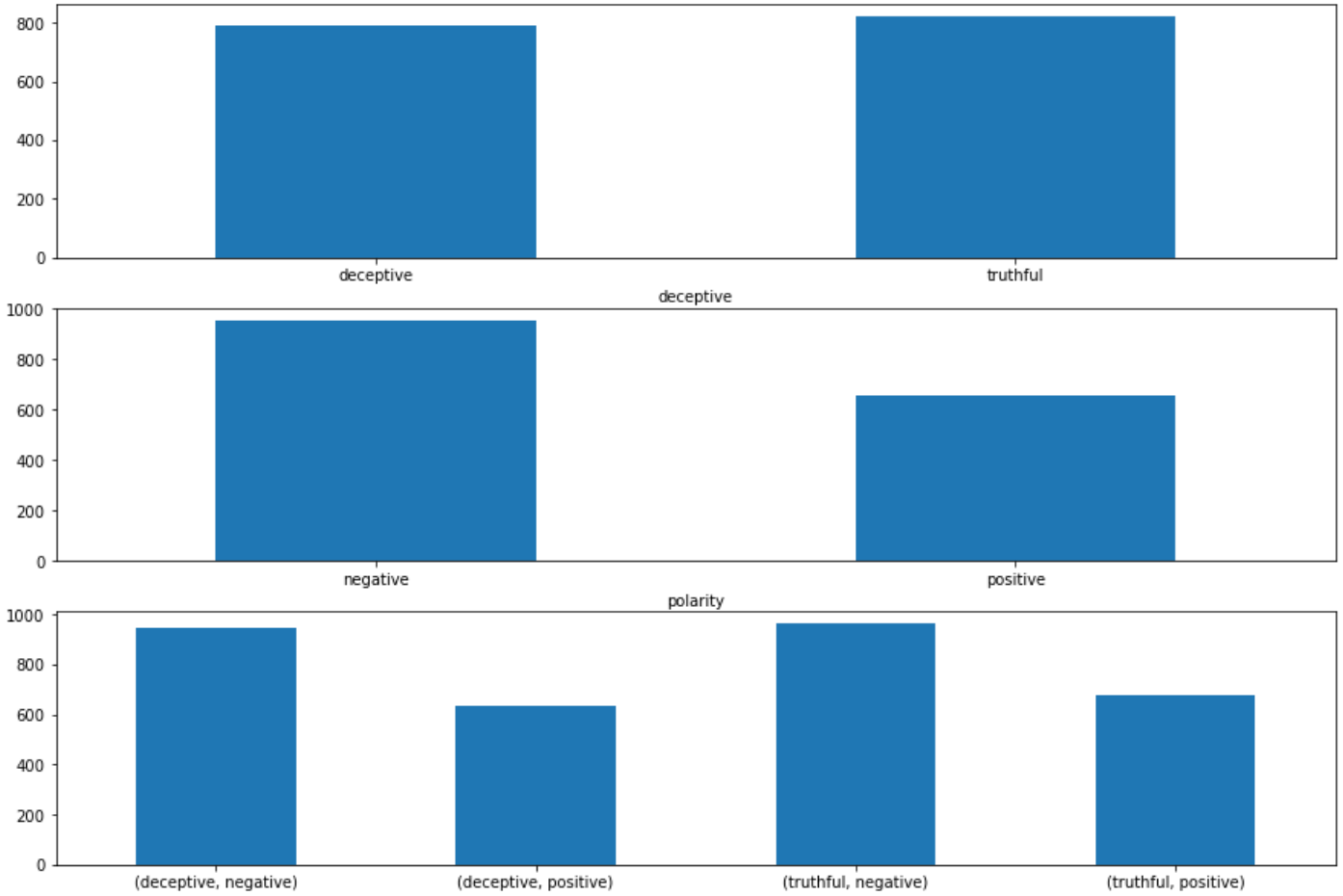
	hotel				text			
deceptive	**deceptive**		**truthful**		**deceptive**		**truthful**	
polarity	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**
source								
MTurk	400.0	400.0	0.0	0.0	400.0	400.0	0.0	0.0
TripAdvisor	0.0	0.0	0.0	400.0	0.0	0.0	0.0	400.0
Web	0.0	0.0	400.0	0.0	0.0	0.0	400.0	0.0

Count of records by Deceptive, Sentiment, and Hotel

	source				text			
deceptive	**deceptive**		**truthful**		**deceptive**		**truthful**	
polarity	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**
hotel								
affinia	20	20	20	20	20	20	20	20
allegro	20	20	20	20	20	20	20	20
amalfi	20	20	20	20	20	20	20	20
ambassador	20	20	20	20	20	20	20	20
conrad	20	20	20	20	20	20	20	20
fairmont	20	20	20	20	20	20	20	20
hardrock	20	20	20	20	20	20	20	20
hilton	20	20	20	20	20	20	20	20
homewood	20	20	20	20	20	20	20	20
hyatt	20	20	20	20	20	20	20	20
intercontinental	20	20	20	20	20	20	20	20
james	20	20	20	20	20	20	20	20
knickerbocker	20	20	20	20	20	20	20	20
monaco	20	20	20	20	20	20	20	20
omni	20	20	20	20	20	20	20	20
palmer	20	20	20	20	20	20	20	20
sheraton	20	20	20	20	20	20	20	20
sofitel	20	20	20	20	20	20	20	20
swissotel	20	20	20	20	20	20	20	20
talbott	20	20	20	20	20	20	20	20

Data (3) Data Exploration

Mean Length of Words Across Categories



From the results of hypotheses tests we can see that:

- There is a statistically significant difference between positive and negative average length of words (p-value of about 0)
- There is a statistically significant difference between trip advisor and non trip advisor sources (p-value of about 0)
- There is a no detectable statistical significance when it comes to deceptive and genuine reviews. (p-value = .2)

Data Preparation: Tokenizing

Sentence → Tokens → Parts of Speech

	deceptive	hotel	polarity	source	text	tokens	tokens_stopwords	lemma	pos
0	truthful	conrad	positive	TripAdvisor	We stayed for a one night getaway with family ...	[We, stayed, for, a, one, night, getaway, with...	[We, stayed, one, night, getaway, family, thur...	[-PRON-, stay, for, a, one, night, getaway, wi...	[PRON, VERB, ADP, DET, NUM, NOUN, NOUN, ADP, N...

Parts of Speech → <u>4 Parts of Speech Variables</u>

pos	pron_ct	noun_ct	punct_ct	verb_ct
[PRON, VERB, ADP, DET, NUM, NOUN, NOUN, ADP, N...	5	30	12	14

Data Preparation: Categorical Data

Categorical Data → Dummy Variables (22 Variables)

hotel	polarity	source
conrad	positive	TripAdvisor
hyatt	positive	TripAdvisor
hyatt	positive	TripAdvisor
omni	positive	TripAdvisor
hyatt	positive	TripAdvisor

hotel_palmer	hotel_sheraton	hotel_sofitel	hotel_swissotel	hotel_talbott	polarity_negative	polarity_positive
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1

Topic Modeling

- Use Latent Semantic Indexing
  - Data Reduction Technique based on “Singular Value Decomposition”
  - Reduces Tokenized Table for each of 1600 review into 300 variables
    - Think of each variable as a “topic” for example:
      - “An amazing Family Vacation” Topic
      - “Honey moon gone awry review” Topic

Classification

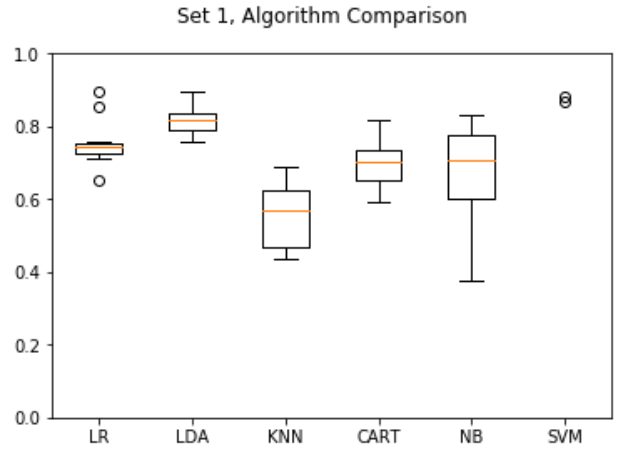
- 6 Classification Models
  - Logistic Regression
  - Linear Discriminant Analysis
  - K Nearest Neighbor Classification
  - Decision Trees
  - Naive Bayes
  - Support Vector Classifier

- Three sets of Features:
  - X1 = Topics Only (300 Variables)
  - X2 = Topics + Parts of Speech Metrics (300 + 4)
  - X3 = Topics + Parts of Speech Metrics + Dummy Variables (300 + 4 + 22)

Results (1)

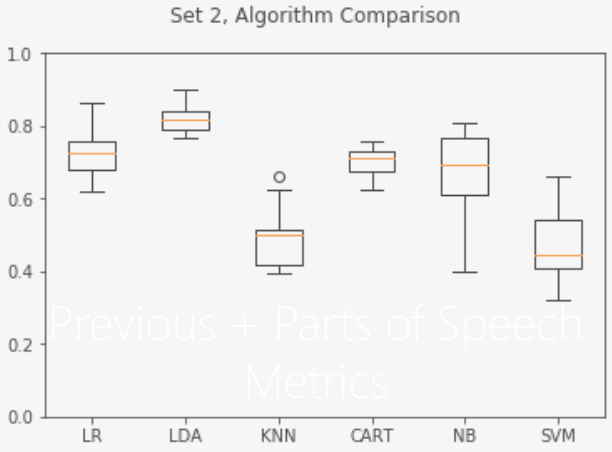
Topics Only

```
LR: 0.755625 (0.066582)
LDA: 0.818750 (0.041174)
KNN: 0.551875 (0.085744)
CART: 0.699375 (0.070459)
NB: 0.665000 (0.148171)
SVM: 0.175000 (0.350011)
```



Previous + Parts of Speech Metrics

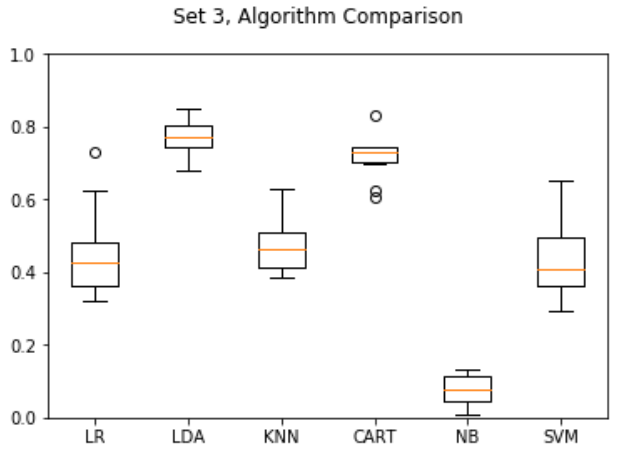
```
LR: 0.727500 (0.070112)
LDA: 0.822500 (0.041477)
KNN: 0.496250 (0.086159)
CART: 0.700625 (0.045022)
NB: 0.661250 (0.133867)
SVM: 0.468750 (0.108253)
```



Feature Set 2 +
LDA performs the best.

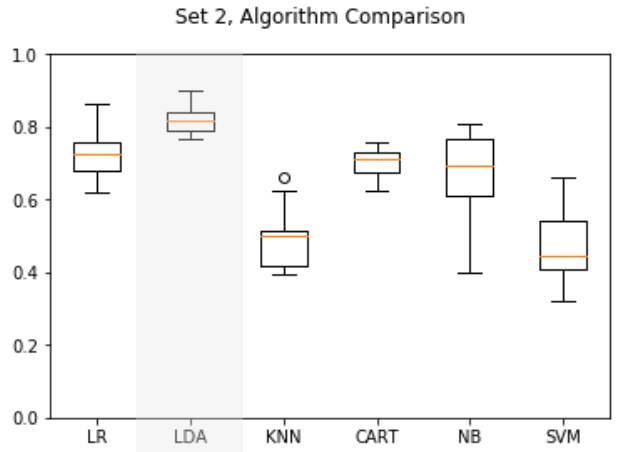
Previous + Dummy Variables

```
LR: 0.456875 (0.125637)
LDA: 0.774375 (0.050191)
KNN: 0.480625 (0.080110)
CART: 0.715625 (0.060418)
NB: 0.075000 (0.041363)
SVM: 0.436875 (0.111329)
```



Results (2)

```
LR: 0.727500 (0.070112)
LDA: 0.822500 (0.041477)
KNN: 0.496250 (0.086159)
CART: 0.700625 (0.045022)
NB: 0.661250 (0.133867)
SVM: 0.468750 (0.108253)
```



Feature Set 2 +
LDA performs the best.

10 Fold Cross Validation Test Results:

Accuracy: 82.25%

Confusion Matrix on Entire Dataset:

	precision	recall	f1-score	support
False	0.94	0.95	0.94	800
True	0.95	0.94	0.94	800
avg / total	0.94	0.94	0.94	1600

Future steps

- The results above suggest that the model performs fairly well under Latent Discriminant Analysis, yet here are some areas of improvement and/or further areas of research.
  - Word misspellings in feature engineering. Train the models with a misspelling indicator, as this could be a feature that might be important.
  - Look at this data across time -- there could be difference in detecting spam based on information around the time period -- for example, will adding month pick up information about holidays and the holiday vacation experience that could help detect fake reviews.
  - Would having transaction data on each review show interesting results -- for example, the time of each post, and location of the IP address that sent the post.